C67 a1

2022-10-02

Question 1A

```
set.seed(1006740421)
a=rnorm(100,mean=100,sd=20)
df1 = data.frame(a)
b=rnorm(1000,mean=100,sd=20)
df2 = data.frame(b)
c=rnorm(10000,mean=100,sd=20)
df3 = data.frame(c)
d=rnorm(100000,mean=100,sd=20)
df4 = data.frame(d)
```

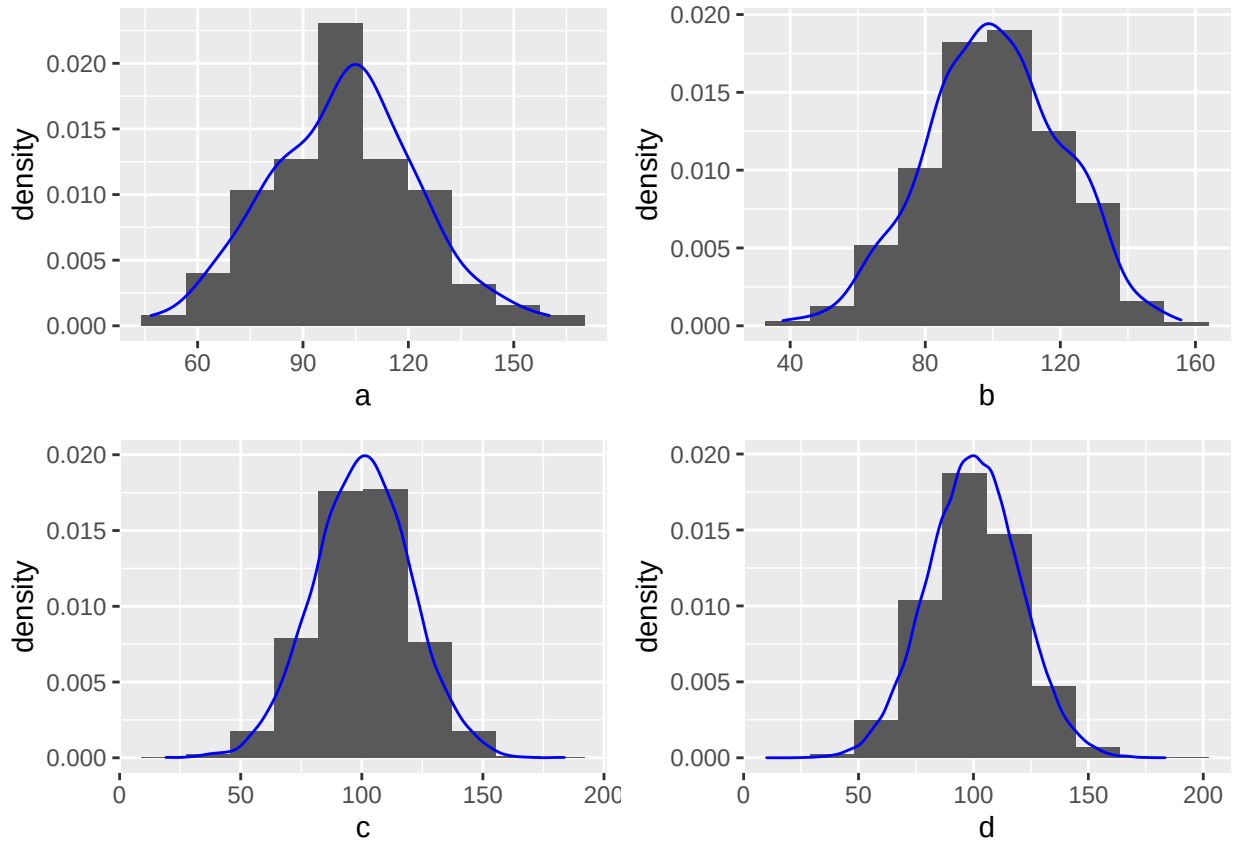
```
p1<-ggplot(df1,aes(a))+geom_histogram(bins=10,aes(y=..density..))+geom_density(col="blue")

p2<-ggplot(df2,aes(b))+geom_histogram(bins=10,aes(y=..density..))+geom_density(col="blue")

p3<-ggplot(df3,aes(c))+geom_histogram(bins=10,aes(y=..density..))+geom_density(col="blue")

p4<-ggplot(df4,aes(d))+geom_histogram(bins=10,aes(y=..density..))+geom_density(col="blue")

grid.arrange(p1,p2,p3,p4,nrow=2)
```



As the size increasing, the data will be closer to a normal distribution.

Question 1B

```
summary(a)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   46.76   87.42  103.07  101.65  114.84  160.14
```

```
sd(a)
```

```
## [1] 20.88165
```

```
quantile(a, c(0.025, 0.25, 0.5, 0.75, 0.975))
```

```
##      2.5%       25%       50%       75%     97.5%
##  63.05151  87.42281 103.07220 114.84038 143.84550
```

```
summary(b)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37.82   86.20   99.87  100.22  113.78  155.75
```

```
sd(b)
```

```
## [1] 20.2446
```

```
quantile(b, c(0.025, 0.25, 0.5, 0.75, 0.975))
```

```
##      2.5%       25%       50%       75%     97.5%
##  60.22967  86.19834  99.87416 113.77874 136.93728
```

```
summary(c)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.15   86.77  100.34  100.24  113.81  183.59
```

```
sd(c)
```

```
## [1] 20.02683
```

```
quantile(c, c(0.025, 0.25, 0.5, 0.75, 0.975))
```

```
##      2.5%       25%       50%       75%     97.5%
##  60.84855  86.77171 100.33805 113.80948 139.55371
```

```
summary(d)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.08   86.54  100.16  100.10  113.51  183.37
```

```
sd(d)
```

```
## [1] 19.99973
```

```
quantile(d, c(0.025, 0.25, 0.5, 0.75, 0.975))
```

```
##     2.5%      25%      50%      75%    97.5%
##  60.9041  86.5363 100.1571 113.5054 139.4848
```

```
qnorm(c(.025, .25, .5, .75, .975),mean=100,sd=20)
```

```
## [1]  60.80072  86.51020 100.00000 113.48980 139.19928
```

For mean, the theoretical value is 100.

Sample size = 100, mean = 101.65.

Sample size = 1000, mean = 100.22.

Sample size = 10000, mean = 100.24.

Sample size = 10000, mean = 100.10.

For sd, the theoretical value is 20.

Sample size = 100, sd = 20.88165.

Sample size = 1000, sd = 20.2446.

Sample size = 10000, sd = 20.02683.

Sample size = 100000, sd = 19.99937.

For percentile (2.5, 25, 50, 75, 97.5), the theoretical value is (60.80072 86.51020 100.00000 113.48980 139.19928).

Sample size = 100, percentile (2.5, 25, 50, 75, 97.5) = (63.05151 87.42281 103.07220 114.84038 143.84550).

Sample size = 1000, percentile (2.5, 25, 50, 75, 97.5) = (60.22967 86.19834 99.87416 113.77874 136.93728).

Sample size = 10000, percentile (2.5, 25, 50, 75, 97.5) = (60.84855 86.77171 100.33805 113.80948 139.55371).

Sample size = 100000, percentile (2.5, 25, 50, 75, 97.5) = (60.9041 86.5363 100.1571 113.5054 139.4848).

We can make a conclusion that as the sample size increasing, the mean, sd, percentile (2.5, 25, 50, 75, 97.5) will closer to the theoretical value.

Question 6

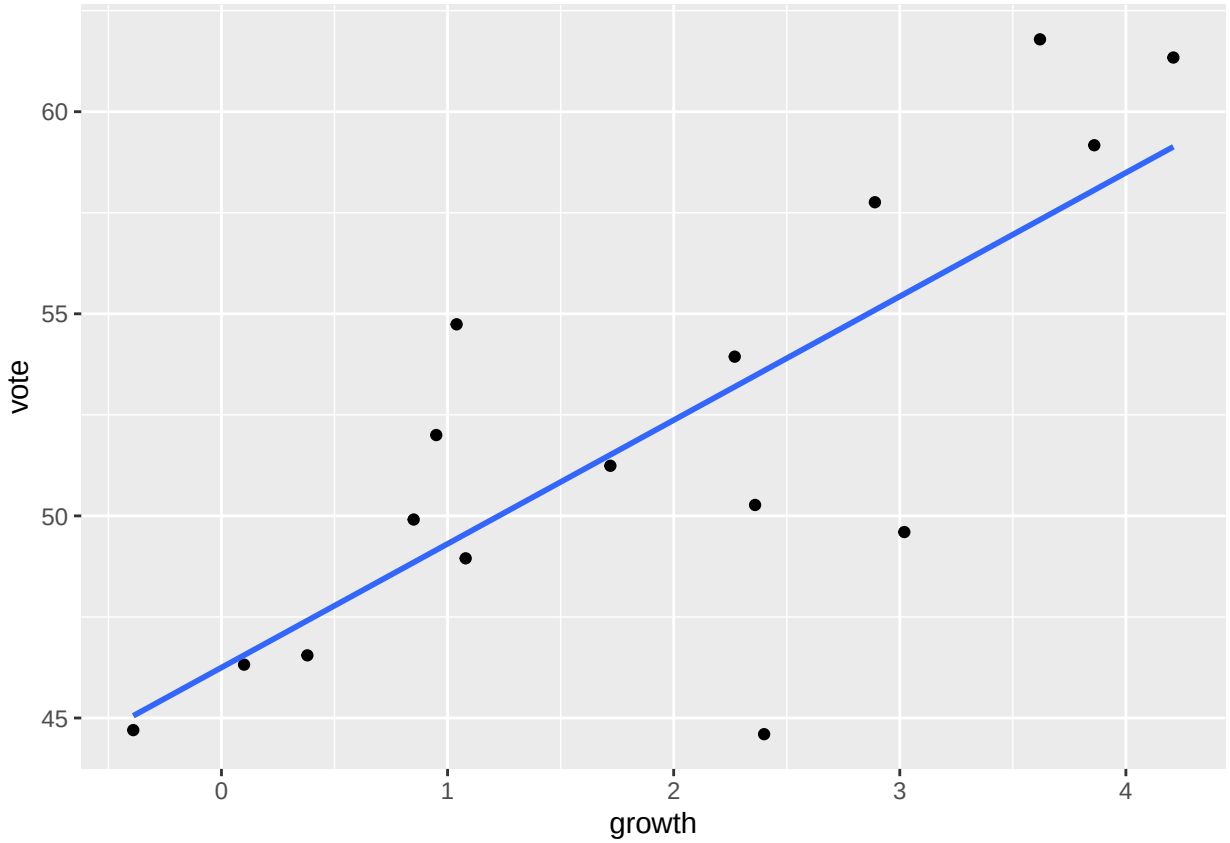
a)

```
data <- read_table('vote.txt')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   year = col_double(),
##   growth = col_double(),
##   vote = col_double(),
##   inc_party_candidate = col_character(),
##   other_candidate = col_character()
## )
```

```
## Warning: 4 parsing failures.
## row col  expected    actual       file
##  10  -- 5 columns 6 columns 'vote.txt'
##  11  -- 5 columns 6 columns 'vote.txt'
##  13  -- 5 columns 6 columns 'vote.txt'
##  14  -- 5 columns 6 columns 'vote.txt'
```

```
ggplot(data,aes(x=growth, y=vote)) + geom_point() + geom_smooth(formula = y ~ x, method=lm, se=FALSE)
```



b)

```
lm <- lm(vote ~ growth , data)
summary(lm)
```

```
##
## Call:
## lm(formula = vote ~ growth, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9929 -0.6674  0.2556  2.3225  5.3094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2476     1.6219  28.514 8.41e-14 ***
## growth        3.0605     0.6963   4.396  0.00061 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.763 on 14 degrees of freedom
## Multiple R-squared:  0.5798, Adjusted R-squared:  0.5498
## F-statistic: 19.32 on 1 and 14 DF,  p-value: 0.00061
```

$\hat{\beta_0} = 46.2476$.It means if the average personal income growth in the pervious year is 0, the incumbent party's share of the two-party vote will be 46.2476 percent.

$\hat{\beta_1} = 3.0605$.It means as the average personal income growth in the pervious year increasing one unit, the incumbent party's share of the two-party vote will increase 3.0605 percent.

c)

In R:

```
predict(lm, data.frame(growth = 0.1))
```

```
##       1
## 46.5537
```

By hands:

```
data
```

```
## # A tibble: 16 x 5
##     year growth  vote inc_party_candidate other_candidate
##    <dbl>  <dbl> <dbl> <chr>               <chr>
##  1  1952   2.4   44.6 "\"Stevenson\""     "\"Eisenhower\""
##  2  1956   2.89  57.8 "\"Eisenhower\""    "\"Stevenson\""
##  3  1960   0.85  49.9 "\"Nixon\""         "\"Kennedy\""
##  4  1964   4.21  61.3 "\"Johnson\""       "\"Goldwater\""
##  5  1968   3.02  49.6 "\"Humphrey\""      "\"Nixon\""
##  6  1972   3.62  61.8 "\"Nixon\""         "\"McGovern\""
##  7  1976   1.08  49.0 "\"Ford\""          "\"Carter\""
##  8  1980  -0.39  44.7 "\"Carter\""        "\"Reagan\""
##  9  1984   3.86  59.2 "\"Reagan\""        "\"Mondale\""
## 10  1988   2.27  53.9 "\"Bush,"           "Sr.\""
## 11  1992   0.38  46.6 "\"Bush,"           "Sr.\""
## 12  1996   1.04  54.7 "\"Clinton\""       "\"Dole\""
## 13  2000   2.36  50.3 "\"Gore\""          "\"Bush,"
## 14  2004   1.72  51.2 "\"Bush,"           "Jr.\""
## 15  2008   0.1   46.3 "\"McCain\""        "\"Obama\""
## 16  2012   0.95  52   "\"Obama\""         "\"Romney\""
```

In 2018, the growth x is 0.1.

$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} * X = 46.2476 + 3.0605 * 0.1 = 46.5536$

"Obama" would win. Because the vote rate for Obama is 54.4464 and the vote rate for Mccain is "46.5536".46.5536 is smaller than 54.4464(1- 46.5536).

d)

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 > 0$

```
summary(lm)
```

```
##
## Call:
## lm(formula = vote ~ growth, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9929 -0.6674  0.2556  2.3225  5.3094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2476     1.6219  28.514 8.41e-14 ***
## growth        3.0605     0.6963   4.396  0.00061 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.763 on 14 degrees of freedom
## Multiple R-squared:  0.5798, Adjusted R-squared:  0.5498
## F-statistic: 19.32 on 1 and 14 DF,  p-value: 0.00061
```

```
1-pt(4.396,14)
```

```
## [1] 0.0003047443
```

Therefore, the test statistics is 4.396.

n = 16, n-2 = 14.

p-value: $P(t_{14} > 4.396) = 0.0003047443$

p-value < 0.05, p-value $< \alpha$. Reject H_0.

There is a positive association between incumbent party's vote share and economical growth.

e)

By hands :

```
qt(0.975, 14)
```

```
## [1] 2.144787
```

95% CI for β_1 : ($\hat{\beta}_1$ - $t_{0.975,n-2} * se(\hat{\beta}_1)$, $\hat{\beta}_1 + t_{0.975,n-2} * se(\hat{\beta}_1)$)

$\hat{\beta}_1 = 3.0605$

$t_{0.975,n-2} = 2.144787$

$se(\hat{\beta}_1) = 0.6963$

3.0605 - 2.144787 * 0.6963 = 1.567085

3.0605 - 2.144787 * 0.6963 = 4.553915

Answer:(1.567085, 4.553915)

By R:

```
confint(lm, level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 42.768951 49.726345
## growth       1.567169  4.553887
```

Answer:(1.567169, 4.553887)

f)

Step 1:depart the absolute sign

$p(|\hat{\beta}_1 - \beta| > 1) = p(\hat{\beta}_1 - \beta > 1) + p(\hat{\beta}_1 - \beta < -1)$

Step 2:multiply both side by

$$\$se(\hat{\beta_1}):\ \$$$

$P(\frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} > \frac{1}{se(\hat{\beta}_1)}) + P((\frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} < -\frac{1}{se(\hat{\beta}_1)})$

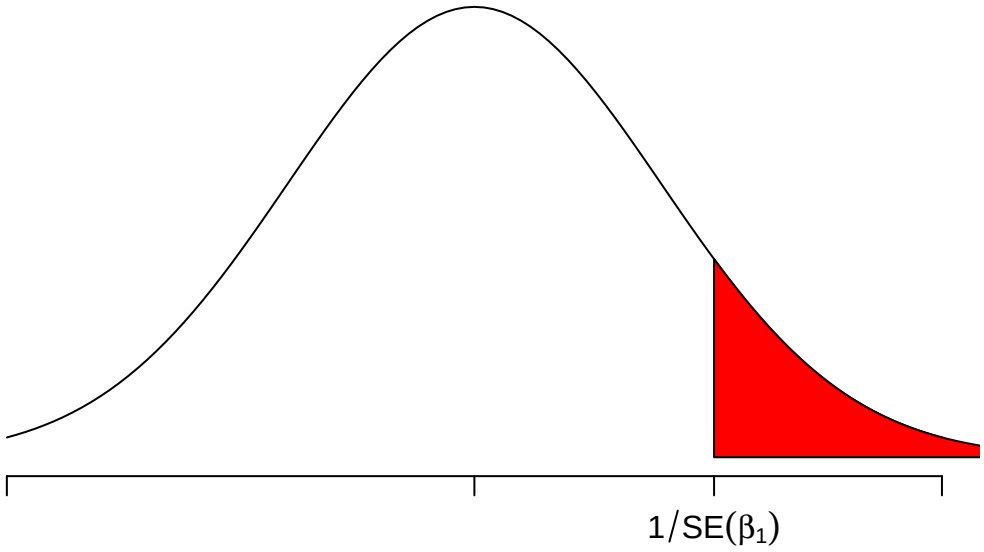
Recall the test-statistic formula and using the porperty to find a more easy-handle data:

$P(\frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} > \frac{1}{se(\hat{\beta}_1)}) = P(t^*(different\ than\ the\ t^*\ showed\ in(b))) > \frac{1}{se(\hat{\beta}_1)})\ P((\frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} < -\frac{1}{se(\hat{\beta}_1)}) = P(t^*(different\ than\ the\ t^*\ showed\ in(b))) < -\frac{1}{se(\hat{\beta}_1)})$

with this data, we can compare to the p-value and using diagram to see the region:

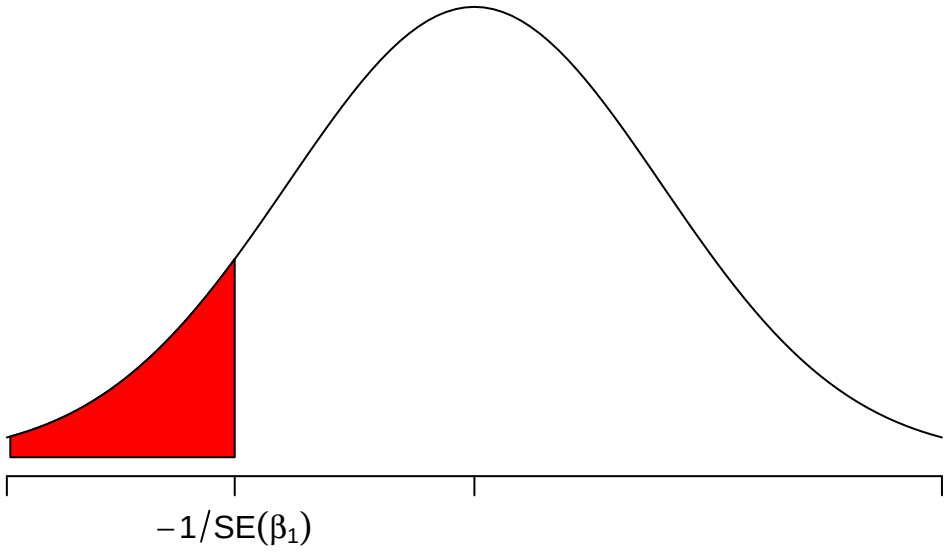
$since\ \frac{1}{se(\hat{\beta}_1)} > 0 -> P(t^* > \frac{1}{se(\hat{\beta}_1)})\ is\ showing\ below$:

```
sD <- 2;
x <- seq(-5,5,by=0.1)
y <- dnorm(x,sd=sD)
right <- 2.5631031310892
tail <- qnorm(0.9,sd=sD)
plot(x,y,type="l",xaxt="n",ylab="",xlab=expression(paste('')),
     axes=FALSE,ylim=c(0,max(y)*1.05),xlim=c(min(x),max(x)),
       frame.plot=FALSE)
axis(1,at=c(-5,right,0,5),labels=c(expression(' '),expression(1/SE(beta[1])),expression(' '),expression
xtail <- seq(right,6,by=0.1)
ytail <- dnorm(xtail,sd=sD)
polygon(c(xtail,xtail[length(xtail)],xtail[1]),
          c(ytail, 0, 0), col='red')
```



since $\frac{1}{se(\hat{\beta_1})} > 0 -> P(t^* > -\frac{1}{se(\hat{\beta_1})})$ *is showing below* :

```
sD <- 2;
x <- seq(-5,5,by=0.1)
y <- dnorm(x,sd=sD)
tail <- qnorm(0.9,sd=sD)
plot(x,y,type="l",xaxt="n",ylab="",xlab=expression(paste('')),
     axes=FALSE,ylim=c(0,max(y)*1.05),xlim=c(min(x),max(x)),
       frame.plot=FALSE)
axis(1,at=c(-5,-right,0,5),labels=c(expression(' '),expression(-1/SE(beta[1])),expression(' '),expressi
xtail <- seq(right,5,by=0.1)
ytail <- dnorm(xtail,sd=sD)
polygon(c(-xtail,-xtail[length(xtail)],-xtail[1]),
          c(ytail, 0, 0), col='red')
```



Based on the comparison and the two diagrams, we can conclude it is just :

```
pt(1/0.6963,14,lower.tail = FALSE)*2
```

```
## [1] 0.1729238
```

Answer: 0.1729238